Chart of Darkness : Mapping a Large Intranet

Peter Bailey, Nick Craswell

Dept. Computer Science, FEIT,
The Australian National University
Canberra ACT 0200 Australia
Peter.Bailey@cs.anu.edu.au, Nick.Craswell@cs.anu.edu.au

David Hawking

CSIRO Mathematics and Information Sciences,
Canberra ACT Australia
David.Hawking@cmis.csiro.au

Abstract:

We introduce and define the concept of *dark matter* on the Web. Dark matter for a person or Web crawler consists of pages that they cannot reach and view, but which another observer can. Dark matter is important to our understanding of the Web in that the portion of the Web any of us can see depends on our viewpoint. Different observers see different overlapping sections of the Web. However, no one can see all of the Web, even if they want to.

We categorise the various types of dark matter that exist and how they may be discovered. Formal definitions of what constitutes lightness and darkness on the Web are formulated in terms of reachability. Our case study of dark matter within the Australian National University's intranet is reported. We estimate that 87% of the ANU intranet's information is dark to our local search service, and 37% is potentially loadable Web data unreachable to almost every Web user. Finally, we discuss some of the implications of dark matter for estimating the size of the Web and for general Web searching.

Keywords: Dark matter; Reachability; Web search; Practice and experience

1 Introduction

The Web has become the pre-eminent starting point for information discovery, and search engines integral to accessing this information resource. Search engines are used by approximately 85% of all Web users to locate information [10].

When searching the Web, users of search engines may be under the erroneous impression that their search will cover "all of the Web." In the last couple of years, researchers have demonstrated that search engine coverage has been declining relative to the size of the Web itself [16]. Most search engine providers place increasing emphasis providing a good *general* coverage of the Web, rather than on providing a comprehensive coverage.

The decline in coverage has been attributed to a number of things, most notably the explosive growth in size of Web servers/pages and the consequent costs of crawling, indexing or providing query processing services over this data. However, there are reasons why complete coverage is not even

possible, let alone practical.

In this paper, we introduce the concept of *dark matter* - information on the Web that is not or cannot be discovered by an individual or a search engine. Our investigations are grounded in a case study of providing search services (`http://search.anu.edu.au`) for a large intranet - the Web servers at the Australian National University (ANU) - which we claim is broadly representative, in many useful ways, of the Web at large.

We outline how Web information is obtained, which leads us to see why certain types of pages are not obtained - the dark matter. The reasons why pages are not obtained leads to a taxonomy of dark matter. We also give formal definitions of dark and light matter on the Web in terms of reachability. A discussion of our experimental methodology for how dark matter may be discovered motivates our experiments in quantifying the prevalence of dark matter in the ANU intranet. Our access to additional information about Web servers within the ANU allows us greater precision in our estimates than would be possible on the Web at large. We also report a simple experiment to reassess the size of the publicly indexable Web. The implications of dark matter for Web search in general and information publishing are also discussed.

2 Related Work

Various people have explored the nature of the Web. Lawrence and Giles' work in estimating the size of the publicly indexable Web [15,16] includes discussion of what can be considered as indexable. Botluk [4] describes some characteristics of what she terms the *Invisible Web*. She is particularly interested in indexes covering documents which are dark to other indexers. A number of companies, including NorthernLight [19] and Intelliseek [12], have constructed directories of searchable databases and collections, facilitating access to documents not obtainable by traditional Web crawlers. Pitkow [20] provides a valuable overview of different characterisations of the Web, although the bulk of these studies explore how people access servers and the rates of change of pages. An analysis of the diameter of the Web when modeled topologically as a graph has been conducted by Albert, Jeong and Barabási in [1]. They estimate the average number of links between any two pages on the Web as nineteen. The Cooperative Association for Internet Data Analysis has been developing tools for exploring and visualising the tomography of the Internet [6]. Their work is concrete in that the live network is probed to obtain information about interconnections, though not specific to Web protocols. We are not aware of any study which has systematically attempted to characterise and measure the quantity of dark matter in the way that we do in this paper.

3 Obtaining Web Information

The main contribution of dark matter as a concept is to characterise and measure why obtaining information is *dependent on the characteristics of the observer*. Observers are most often either people who download pages from the Web using some browser technology, or crawler agents of certain kinds of Web search systems. There are also other automatic Web data downloading systems, such as those used for mirroring archives.

People typically find pages by: learning of the existence of a URL, following links from pages which they download, forging URLs based on cultural knowledge, or using search systems which present them with pages of links. Forging a URL involves constructing a URL based on your information needs (eg. `news.com` for news information), or modifying some existing URL.

Current Web search systems fall into three main categories, based on how the data is gathered. There are directory-based systems, which rely on editorial selection, categorisation and description of useful sites. The second approach is based on using indexes constructed from Web pages that have been collected by a Web crawler. Popular examples include AltaVista [2] and Google [9]. The third approach is distributed information retrieval or meta-search systems, which probe other search systems to find results. In this paper, we are particularly interested in search systems which use Web crawling to obtain Web pages for indexing, since these are most directly affected by the existence of dark matter.

The basic architecture of a search engine such as Google which uses Web crawling to provide its raw data is described well in [5]. The search service we have engineered at the ANU follows a similar construction, with crawler, indexer and query processing components. The component which actually obtains the raw data for indexing is the Web crawler. Crawlers are also known as robots or spiders. Simply put, a Web crawler behaves as follows: given some starting set of URLs, the corresponding contents of the pages are retrieved by the crawler, and scanned for new URLs. Some policy encoded in the crawler is used to decide whether each new URL should be added to a queue of URLs to be retrieved later. A simple policy is to check whether the URL has already been loaded, and to ignore it if it has. The downloaded contents of each URL may be stored locally, or discarded once the indexer component has examined them.

4 Dark Matter In WebSpace

The term "dark matter" was coined by astronomers to describe observations by Fritz Zwicky over 50 years ago that demonstrated a substantial amount of the matter of the universe is not visible; in technical terms it is non-luminous. Over the course of the last 50 years, estimates by astronomers of the amount of dark matter in the universe have increased to more than 90%.

4.1 Why Is There Dark Matter?

Dark matter seems an equally appropriate term for an analogous situation in Web space. We understand the Web to be all pages which can be accessed through the HTTP protocol by some person or computer somewhere on the Internet. Thus a page which has no incoming links and no outgoing links is still part of the Web provided someone knows its URL and can gain access to its content from some suitable Internet IP address. Note, we believe that the identity of a page is not solely its URL, but is characterised by its content as well. We can "see" parts of the Web using search engines, which act as our telescopes, collecting and organising information for us. But there are parts of the Web which are not visible from any of these search engines.

In 1997 [21], it emerged to public surprise that the AltaVista search service [2] was not indexing the entire Web. Since then, estimates of the size of the Web have been made based on probing commercial Web search engines and checking for overlap [3,15,16]. Lawrence and Giles's later study estimated that in February 1999 the Web contained over 800 million "publicly indexable" pages. They also found that the biggest search engine at that time, Northern Light [19], indexed approximately 16% of the available Web. They estimate that there can often be substantially less than 50% overlap among search servers. Combined coverage of all the search engines examined was approximately 42% of the available Web. These proportions are percentages of the pages visible to search engines.

However, it is not just a matter of having inadequate resources to crawl the Web. There is much of the Web which is completely invisible to the Web crawler agents of these search engines. In a response to the issue of AltaVista's incomplete coverage of the Web in 1997, Louis Monier, then Chief Technical

Officer at AltaVista, made the following comment (as reported in [21]):

"The concept of `the size of the Web' in itself is flawed, as there are many sites virtually infinite in size: dynamically generated documents, personalized news pages and shopping baskets using cookies, robot traps, scripts, the list goes on."

Monier's remarks strike at the heart of the dark matter issue - there are many reasons why search engines in particular (and also human browsers) are unable to find material that exists on the Web. Some of these are technical limitations, and some are due to the nature of the Web itself. In particular, there are many significant and valuable sources of information which are dark matter, not just Web shopping baskets.

4.2 Shades Of Darkness

There are various shades of dark matter on the Web. The reasons why material is not discoverable give rise to a classification taxonomy, summarised in figure 1.

Rejected	Restricted	Undiscovered	Removed (temporal)
Did not load	Could not load	Could not find	No longer available
Resource limit	Password	Linking pages dark	URL has new content
File type	Domain	or unparseable	No longer any page at URL
Exclusion	Intranet	No linking pages	
CGI			

Figure 1: A taxonomy of dark matter and the reasons for its existence.

4.2.1 Rejected Dark Matter

The first kind of dark matter arises because the observer *rejects* a page for some reason and thus does not load it. Human observers reject pages due to resource limits or if the page requires a particular kind of browser plug-in software that they do not have. Web crawlers frequently reject pages for policy reasons. For example, most current search engines' Web crawlers have finite limits to the number of documents crawled per server. Web crawlers may also choose to exclude themselves from crawling parts of a site by following the `robots.txt` convention [13] or meta tags in the page contents. The page loading policy of a Web crawler may also exclude files with particular extensions. For example, the crawler of our search service does not attempt to load files with `.tar.gz`, `.jpeg`, or `.gif` extensions and many others.

An interesting subcategory is material which is generated dynamically. We refer to this as *generated dark matter*. For example, many pages are generated by programs. Some of the pages are assembled from database information, or constructed based on data passed in forms. Web crawlers typically do not exhaustively probe these areas because it it is too computationally expensive to do so and/or the task of filling in the forms appropriately cannot be automated. Generated pages within a site may in fact be infinite in number.

Since crawlers use rule-based exclusion, they typically rely on some indication within the URL that the page is generated. A URL which contains `cgi-bin` as the base of the URL path or the presence of ?'s are common indications. If such markers are not present, then the pages are likely to be crawled because they are undetectable as generated information. Obviously, generated content is rarely dark to human observers because we do not limit our browsing based on the format of a URL.

4.2.2 Restricted Dark Matter

Second, there is material which is publicly linked to, but is *restricted* to observers with the appropriate permissions. Without these permissions, the pages cannot be loaded. For example, many of the pages for SUN's Java Developer Connection resources [22] are inaccessible without filling in a form to generate a user/password combination first. This user/password combination is a common mechanism used by Web sites to restrict access to specified users. Without such user/password combinations, Web crawlers cannot access the protected information, even if creation of the user/password combination is free.

Similarly, domain restrictions can be used by a Web server administrator to limit access to particular content. The contents of the primary Samba software server (`samba.anu.edu.au`) are restricted to local Australian access only, to avoid large international network traffic costs which should go to the US mirror sites for Samba instead. Domain restrictions are also the essence of network firewalls, which prevent observers from outside the firewall from accessing information located within the firewall. For observers within it, the information is not dark to them.

4.2.3 Undiscovered Dark Matter

The third kind of dark matter encompasses pages which remain *undiscovered* by an observer. The reason for this is straightforward: the necessary links which locate the material are never found, being in pages which are dark to the observer. Hence any pages to which these links refer do not get discovered (unless the links are also found in pages which are not dark matter). Another reason why links may not be found is that they exist in loaded pages, but there is no ability to extract them from the information. For example, links in Adobe PDF documents are rarely extracted by Web crawler software.

An interesting subcategory of undiscovered dark matter is material which is not publicly linked to at all. We refer to such pages as *private dark matter*. Access to the material relies on knowing or guessing the URL. For example, information can be placed on a Web server and the URL sent to someone via email, who can then download it. Of course, this person might then put the URL on one of their public pages, and then it would no longer be private dark matter.

The set of starting URLs is also critical in determining what is undiscovered dark matter. When we began our current study of the ANU intranet, pages about AusDance, the Australian Dance organisation, which are hosted on `sunsite.anu.edu.au/ausdance`, were not discoverable from our search service. From this we were able to infer that there was no page containing a link to `sunsite.anu.edu.au/ausdance/*` reachable from our starting set of host URLs (which includes `sunsite.anu.edu.au`) according to our loading policy. However, the AusDance pages were discovered by searching for `ausdance AND host:anu.edu.au` on the AltaVista service. From this we can infer that there were pages containing links to the AusDance site, but these links probably existed externally to the `anu.edu.au` domain, or that they added their URL manually to AltaVista's starting set. (Since the start of our study, the ANU intranet has evolved, and there are now links to the AusDance site within the intranet that we discover, and hence it is no longer undiscovered dark matter to our search service.)

4.3 Web Pages Can Be Dark In Multiple Ways

As discussed above, Web matter can be dark for different reasons. Any single page may be dark for more than one reason as well. What follows is an example exhibiting multiple shades of darkness.

For example, while developing the search service used as part of these experiments, we made it available on the ANU intranet. However, we did not make any public links to it. Hence, it was *undiscovered private* dark matter to anyone to whom we did not pass on the URL.

We did not want any part of the server to be crawled, so we also installed a `robots.txt` exclusion in case anyone to whom we had passed the URL also made links to our service from publicly accessible pages. Hence it was *rejected* dark matter to commercial Web search engines which adhere to the standard for robot exclusion.

Much of the material provided is generated by a Perl CGI script, so even if a search engine did discover it, most pages would still be *rejected generated* dark matter.

Due to privacy issues, since some of the material we crawled was restricted to ANU-only access (accessible to us because we crawled from within the `anu.edu.au` domain), we restricted access to the server to be only within the ANU intranet. To an observer within the ANU its pages were light, but otherwise it was *restricted* dark matter.

4.4 A Definition Of Darkness

The main difficulty in defining what constitutes dark matter on the Web is that darkness is relative to who or what constitutes the observer. A secondary difficulty in defining dark matter is that any observation of the Web is highly dependent on the time at which it is made. Worse still is that in any experimental observation of the Web, the observation must be carried out over a period of time, not a single instant. The effect of this is similar to Heisenberg's Uncertainty Principle: the more precisely we can determine what is dark on the Web to some observer, the less precise we can be about the time at which this information held true.

If we abstract from specific types of dark matter, we see that there is really only one reason why matter is dark: the matter cannot be reached by a particular observer. This realisation leads to the following definitions for characterising darkness. For the purposes of these definitions, without loss of generality we make the simplifying assumption that the entire Web is instantaneously observable at some time t. We return to temporal issues in section 4.6.

First we must define precisely what we mean by matter being reachable by an observer.

Definition - One-Step Reachability. An observer o is defined by their location at some IP address m, and their loading strategy l, written $^{o}\langle m,l\rangle$. The page contents p_u of a URL u are *one-step reachable* with respect to an observer $^{o}\langle m,l\rangle$ if, given the page contents of a seed set of URLs p_S (S is the set of URLs), then: u can be extracted from the page contents of p_S, u is chosen to be loaded according to l, and the actual page contents p_u are successfully loaded by o. We consider p_S to be reachable from p_S by default since they are already known. In practice, $o$obtains the page

contents p_S by successfully loading all URLs in *S*.

The loading strategy *l* is constrained by the following:

- **o's location, *m*, which bears an association with some Internet address**
- **the set of access permissions that *o* possesses, such as passwords**
- **a link extraction policy, *ep*, which determines the file formats which can be parsed to discover new links**
- **a page loading policy, *lp*, which determines which URLs are desirable to load and, of those, which will be chosen for loading**
- **the capabilities available to initiate link generation from visited pages (eg. generating search requests from search server pages)**

The page loading policy encompasses decisions such as:

- **whether to ignore robot excluded pages**
- **whether to look at `cgi-bin` and other detectably generated pages**
- **limits to loading pages: limits to certain sites or domains,limits in discovery depth, etc.**

Definition - Reachability. The foundation of *reachability* is one-step reachability, a single successful page load of a URL which we extract from the page contents already reached. We are interested in defining sets of page contents that are reachable, either in a single step or in several steps, by an observer given the URLs extractable from our initial set of page contents.

Given the entire set of pages on the Web, which we denote *WEB*, we define a *reachability relation*, $R(\ ^{o\langle m,l\rangle}\ )$, from p_S to a set of pages $p_S \subseteq WEB$. As noted before, $p_S \subseteq p_S$.

The reachability relation is transitive as follows. If $p_S\, R(\ ^{o\langle m,l\rangle}\ )\, p_{S'}$ and $p_{S'}\, R(\ ^{o\langle m,l\rangle}\ )\, p_{S''}$ then $p_S\ R(^{o\langle m,l\rangle}\ )\, p_{S''}$. Remember that $p_S \subseteq p_{S''}$.

The page contents p_u of a URL *u* are *reachable* for an observer $^{o\langle m,l\rangle}$ if $p_u \in p_S R(\ ^{o\langle m,l\rangle})^*$, the transitive closure of $R(o\langle m,l\rangle)$ with respect to the page contents of the seed set of URLs p_S.

A consequence of the definition of reachability is to be able to define which pages are light for a particular observer.

Definition - Lightness. The set of pages, *LIGHT*, which are *light* to an observer $o\langle m,l\rangle$ is the set of pages which are *reachable* from the page contents of a seed set of URLs p_S. Thus,

$$p_S R(\ ^{o\langle m,l\rangle}\ )^* \equiv LIGHT.$$

Similarly, we can define whether the page contents of some URL are dark to an observer, and the entire set of dark pages.

Definition - Darkness. The page contents p_u of a URL u are *dark* with respect to an observer $o_{\langle m,l \rangle}$ and given the page contents of a seed set of URLs p_S if $p_u \notin p_S R(o_{\langle m,l \rangle})^*$.

We may also define the set of dark pages *DARK* for $o_{\langle m,l \rangle}$ given the page contents of a seed set of URLS p_S as $WEB \setminus LIGHT$.

A significant aspect of our definitions is that simply because the URL of a page is known, does not make its contents light to an observer. The contents of the page must also be loadable for the page to be light.

It is especially important to recognise that darkness is a relative concept. Just because a page is dark to one observer, does not mean it is also dark to another. For example, someone with a particular user/password combination can see pages which someone who does not have that combination cannot see. The same observer may see different sections of the Web depending on where they access it; from within a firewall at work, but outside the firewall at home. Similarly, a Web crawler may choose to obey a robot restriction for part of a site, whereas a person does not obey this restriction, and is unlikely even to be aware of it. These parts of the site are dark to the crawler, but light to the person.

Another point to consider is that just because a page is reachable, does not mean that its links can be parsed. For example, a GIF image may be reachable by a search engine, but it is unlikely that if the image contains a graphical representation of a URL, then this URL will be understood and used to find new links. As humans, we may read the link with ease, type it into our browser and follow it to its destination. Naturally, reachable pages may be indexed, regardless of whether they can be parsed for link extraction, as evidenced by image search software.

Typically a file format which is parseable (for indexing) is also one from which links can be extracted. However, this is not always the case. Consider a text file which contains `www.anu.edu.au` and `D.A.Hawking`. These words are parseable, but the first is not necessarily identifiable as a link which should be extracted, and the second is not a link.

4.5 Publicly Indexable Web

A consequence of these definitions is that we can define more precisely what is usually meant by the "publicly indexable" Web referred to by Lawrence and Giles [15]. They explicitly discount servers which require authorisation of some kind - what we classify as restricted dark matter. They also remove such things as Web-hosting companies which present a home page on multiple IP addresses (keeping a single representative example), and printers/routers etc which have Web addresses but do not contain significant content. They do this to gain a better estimate of how much real information is available on the publicly indexable Web; they are not attempting to formally define it. In terms of our definitions, however, the page contents of any server which is reachable without restriction should be included. In terms of our definitions, our intuition is that the publicly indexable Web consists of those pages in common which are *light* for many observers. Since every observer is likely to have a small fraction of the Web that they can reach, but which is restricted (typically by domain permissions) from general access, we need to find a way to eliminate these areas which are dark to others. For example, the ANU's intranet contains pages which are not accessible to observers outside the ANU, and hence these pages are not light to everyone.

Definition - Publicly Indexable Web. Let $WEB \equiv p_W$ and construct a page p_a containing links to all $u \in W$. Hence all possible URLs to pages which are to be found in *WEB* are known. Choose an observer *o*, with a loading strategy *l* which has no special access permissions or capabilities. The page loading policy *lp* is chosen to select only *indexable* file formats, and will load all such pages it can discover. Note that the link extraction policy *ep* is immaterial because we have constructed a page from which all possible pages are one-step reachable.

Then the *publicly indexable Web* is a set of pages $INDEXABLE \subseteq WEB$ such that

$INDEXABLE \equiv \bigcap_{p_a} R(\ ^{o(m_i,l)})^*$ for all $i \in M$, where *M* is a *suitable* set of random locations.

The choice of *indexable* file formats is dependent on the observer. This choice will determine the nature of the publicly indexable Web that they see.

The choice of a *suitable M* is of course critical. No $m \in M$ should be chosen such that *m* is restricted from general access to the Web at large, otherwise on forming our intersection set, *INDEXABLE* will be $\emptyset$. Similarly, *M* should be a broad enough set to include Internet domains which are denied access by some Web servers. For example, most of `samba.anu.edu.au` denies access to people accessing from outside Australia. Having a location in *M* which was not in Australia would cause it to be dark for that location, and thus it would not be included as part of the publicly indexable Web.

In figure 2, we show the relationship between various parts of the Web. The publicly indexable Web under our definition is a superset of the pages intended by Lawrence and Giles in their discussion of the publicly indexable Web. Both are smaller than the Web as a whole, due to all the dark matter, and both are larger than any existing search engine's coverage of the Web.

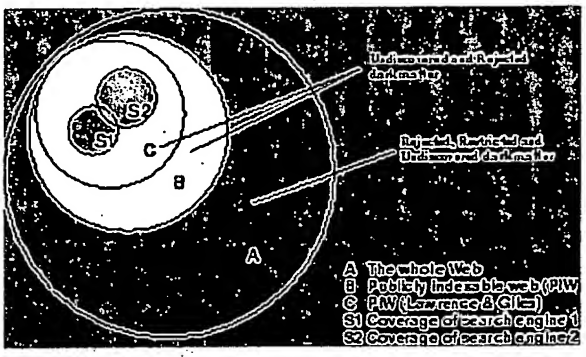

Figure 2: Different parts of the Web.

4.6 Removed Dark Matter

The Web changes constantly: new Web servers go online, old servers are rebooted or turned off, new pages are created, old pages are changed, or removed, and intermittent network and server failures prevent access to entire areas of the Web. Because we consider the page contents of a URL, not the URL

alone, an additional category of dark matter arises as a consequence of the volatility of light matter when considered in a temporal context.

This fourth kind of dark matter is that which is available at some point in time, but later on has been *removed* and is no longer available. There are two subcategories of removed dark matter, which are differentiated by whether the URL is still valid, or whether it too has vanished.

4.6.1 Ephemeral Dark Matter

Removed *ephemeral dark matter* exists when information which was once available is removed, and new information takes its place. For example, the discussions that take place in HTTP MUDs and online-chat rooms are rarely archived, and thus never discovered by a Web crawler. Similarly, many media organisations replace the entire contents of their news pages on a daily or faster basis.

More formally, the page contents p_u of a URL $u$change, although u does not. We can capture this by adding our notion of observation at a particular time to our notation for page contents. We write $p_{\langle u,t \rangle}$ to indicate the page contents of a URL u that were successfully loaded at some time t. Then $p_{\langle u,t \rangle}$ is ephemeral dark matter if $p_{\langle u,t \rangle} \neq p_{\langle u,t' \rangle}$ where $t \prec t'$ for some observer $o_{\langle m,l \rangle}$

4.6.2 Dead Dark Matter

The other subcategory of interest is *dead dark matter*. Dead dark matter is information which has been successfully loaded at time t, but when an attempt is made to observe it at time t', it is no longer available. Sometimes its prior existence is indicated within a search engine's index. Estimates of dead links in the various search engines' indexes ranged from 2.2% to 14% in Lawrence and Giles's study [16]. In general, all Web matter is likely to become dead dark matter. It just depends on how long the time period is between observations.

Again we use our expanded notation of page contents to define dead dark matter more precisely. If $p_{\langle u,t \rangle} \in LIGHT$ for some observer $o_{\langle m,l \rangle}$ with starting set $p_{\langle S,t \rangle}$, and if $p_{\langle u,t' \rangle} \in DARK$ where $t \prec t'$, then $p_{\langle u,t' \rangle}$ is dead dark matter (at time t'). Such a scenario usually occurs because u is either no longer discoverable, or if it is, the page contents $p_{\langle u,t' \rangle}$ are no longer reachable and u is now a dead link.

4.6.3 Observation Within Time Periods

The notion that we can instantaneously observe the entire Web is clearly laughable. If we could view a page every second, it would take more than 25 years to observe the 800 million pages currently estimated [16] as the size of the publicly indexable Web. To observe all these pages in a minute we would need to view 13.3 million pages every second.

A more realistic portrayal of observation might be to replace our notion of an instantaneous point in time t, with a period of time $T \equiv t..t'$. Again, we can replace our existing use of t with Tin the definitions given earlier without loss of generality. The notion of an observation carried out within some

time period more closely correlates with experimental methodology in carrying out observations of the Web. In particular, the action of successfully loading page contents of some URL takes some finite period of time. Loading many pages takes even longer.

Frustratingly, it would be possible given an appropriate time period for page contents to be both light and dark under such definitions. We suggest that page contents should be considered light if there exists some time $t_\alpha \in T$ such that $p_{(u,\alpha)} \in p_{(S,T)}R(o_{(m,l)})^*$. The reason for choosing lightness rather than darkness as the default is that in real experiments we have more information about the pages we are able to load. We do not know in general if there existed some time during our experiments when the same page was not reachable.

5 Experimental Methodology

5.1 The Servers

The Web servers at the Australian National University provide the basis for a case study of dark matter on the Web. There were at least 174 individual HTTP servers whose Web addresses lie within the `anu.edu.au` Internet domain during the week we conducted our experiments. In combination, the servers act as the ANU intranet. (We do not consider other types of servers, such as FTP servers, in this paper.)

There is little central control imposed on these servers, and there is a wide variance in the level of conformance to any kind of standard. There are differences in the server software used, the underlying hardware, and the size of the servers' collections. A huge variety of information is represented within this intranet, including for example people's home pages, department information, software archives, newsletters, research papers, course information and so on. Web material has been available from servers at the ANU since at least 1993.

Thus the ANU intranet acts as a reasonable sample of the Web at large. Although not a random selection of Web servers, the advantage of obtaining extra information from these servers (through the server administrators) is a key enabling feature in conducting this study. For example, we can both crawl the ANU intranet's Web servers and obtain actual directory listings of the Web data to identify dark pages. The ANU is also an example of a large organisation where much of the information is both part of an intranet, and simultaneously part of the Web.

5.2 Technology Used

The basic technology we used for crawling the ANU intranet was a modified version of the GNU `wget` software [18]. The primary modifications to `wget` enable multi-threaded retrieval. The `wget` software adheres to the `robots.txt` convention for Web crawling [13]. It also allows for delays between consecutive requests. Delays between requests to a particular server are especially important for ethical behavior of a multi-threaded crawler. Otherwise, individual servers can be overloaded with multiple simultaneous requests.

5.3 What Is Out There

As with all things on the Web, the nature of the data is highly dependent on the date on which the data was collected. The data we used for these experiments was obtained from the ANU intranet in the week

commencing 15th October 1999.

A total of 174 servers provided pages which we were able to download. The total number of pages downloaded was 292192, and thus the average number of pages per server was 1681. (As is usual with Web servers, the distribution of pages is skewed across servers, approximating a universal power law [11].) There were 69521 unique links in these pages which referred to servers outside the ANU intranet domain.

Dark matter is difficult to discover, let alone quantify. With the assistance of various Web server administrators, we obtained directory listings from 28 of these servers. The directory listings were produced by using a UNIX `find . -type f` command in the server's HTTP directory root. In some cases, this mechanism included `cgi-bin` and image directories, but depending on the setup of servers, it did not always cover all areas. For example, users' home page directories were not always included. We are only able to perform analysis on the directories which were made available.

With additional information, it would be possible to measure more about the presence and quantity of dark matter. For example, Web server access logs can be used to determine how many requests for dark matter pages have been made. Alternatively, if it is feasible to packet sniff the network it is possible to log which requests are for data which is accessible on the Web, but not known about currently. In general however, packet sniffing on networks is not a technology available to the average dark matter researcher. In any case, a sniffer can only examine the packets which pass a particular point in the network. We did not use packet sniffing or Web server access logs in this study.

5.4 The Documents

Ideally, we would have directory listings (and HTTP access logs) for every server that we crawl, but this is not the case.

The crawler's log gives us one way of discovering some dark matter. The estimates we can produce with this data are fairly coarse, since we know very little about pages we do not load other than their possible availability to some Web user. These estimates only relate to the rejected and restricted dark matter categories. We analyse why our crawler does not load particular kinds of files or what kind of restriction was encountered.

The directory listings enable us to provide additional estimates of the real amount of dark matter in some categories, though only for parts of 28 servers. In particular, since we know the loading policy of our crawler, we are able to estimate reachability for three scenarios based on varying the seed set of URLs.

6 Mapping An Intranet's Dark Matter

The breakdown of light and dark matter is shown in table 1. The dark matter is split into the three categories we can detect from the crawler logs alone. Undiscovered dark matter cannot be detected from crawler logs and nor can removed dark matter. The second column contains the number of URLs in each category from our crawler's log data.

Table 1: Quantities of light matter and dark matter by category for the ANU intranet.

Category	Data from Web crawler log: number of pages	Data from directory listings: mean % of pages on a server
light	292192	50.2%
rejected	768500	49.3%
restricted	1178	0.5%

From the set of servers for which we have directory listings, we report the average percentages in each category compared to all files. These are summarised in the third column of table 1.

6.1 Mapping the Rejected Dark Matter

As stated in table 1, our ANU intranet crawl found 768500 URLs which were rejected. There are a number of reasons for their rejection.

The generated dark matter forms an interesting subcategory. This result for generated dark matter counts all links which contained a reference to a `cgi-bin` filename stem and all URLs which were found with `?`'s contained in them (removing multiple occurrences which had the same base prior to the `?`). It is assumed that at least one page for each such link is generated. There is a significant skew across the servers in that one server has more than 10000 generators (in fact, just over 27000), two servers have between 1000 and 10000 generators, five servers have between 100 and 1000, thirteen have more than 10 and less than 100, and fifty have less than 10.

Another interesting subcategory is those rejected due to robot exclusion, although only 39 of the 174 servers actually contain `robots.txt` files of any description.

The remainder are those which are rejected on the grounds that our crawler cannot extract links from their file formats. We choose to break these down into 5 major areas: image files - *images* (`gif`, `jpeg`, etc); readable document files - *text* (`pdf`, `ps`, `doc`, etc); program code files - *prog* (`c`, `cc`, etc); backup files - *backup* (, `.bak`, etc); and everything else - *other*. (We could give more detailed breakdowns, but there are far too many unique extensions for this to be meaningful.)

The breakdown for each rejection reason is given in table 2. The percentage of URLs in each category, relative to the total number of rejections, from our crawler's log data is given in column two of the table.

Table 2: Quantiles of rejected dark matter by kind.

		Directory listings data:

Reason	Web crawler log data: % of rejected pages	mean % of the rejected pages on a server
robot exclusion	0.7%	5.5%
generated	3.6%	1.0%
text unindexable	2.8%	8.8%
image	12.1%	54.3%
prog	3.9%	8.9%
backup	0%	1.8%
other	76.9%	22.8%

When examining the server directory listings, we have more precise estimates of the quantity of rejected dark matter. These are summarised in column three of table 2. The results are presented as average percentages of the total rejections of a server's files, based on the listings from 28 individual Web servers. We note that as described earlier for the rejected generated dark matter, typically there is significant skew in the results, with some servers having many files rejected for a particular reason, and others having none.

The percentage of rejected files of text format kind (for which in general there exist publicly available filters which convert to text) suggest that we should be downloading these pages as well. If the percentage extrapolates to the ANU intranet as a whole, there would be an additional 104000 documents that could be indexed by our search service.

6.2 Mapping the Restricted Dark Matter

The ANU intranet crawl logs are able to give us precise reasons as to why files are not loadable. The HTTP error codes for authorisation required and access denied are reported separately. Of the 1178 restricted URLs, 729 were due to access restrictions and 449 required user/password authorisation.

We used the longest directory paths from these restricted URLs, and mapped them onto the server directory listings to estimate how many files might be restricted on individual servers. (There is no actual way to determine this however, without actually trying to load these files.) Of the 1.31% average of a server's pages which are restricted from loading, the access restricted areas per server constitute 1.09% and the authorisation required areas per server is a bare 0.22%.

6.3 Mapping the Undiscovered Dark Matter

The crawler's log does not permit any calculation of the amount of undiscovered dark matter. Instead, we look at the differences between what the crawler loads, and what is theoretically loadable were we to have access to all URLs that exist on the servers for which we have directory listings. We also include the amount loadable when we crawl with just the server's base URL as the seed set and restrict discovery of pages to the server alone. Given the current loading strategy, figure 3 demonstrates how different seed sets of URLs affect the quantity of Web pages which are light to an observer.

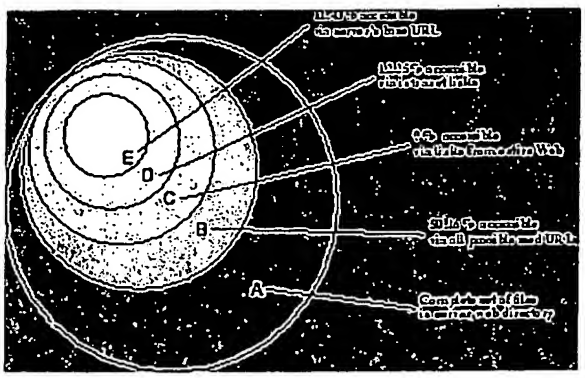

Figure 3: What is light depends on what you know about to begin with according to our crawler's loading strategy. Average percentages of pages from the set of 28 server directory listings.

The difference between what we actually load in our crawl of the ANU intranet (an average of 12.25% of all pages on a server) and the theoretically loadable pages were we to have all their URLs (an average of 50.16% of all pages on a server) gives us the amount of undiscovered dark matter. The difference is 37.9%, which if extrapolated to the whole of the ANU intranet, allows us to estimate that as many as 903000 pages may exist as undiscovered dark matter! In other words, the roughly 292000 pages we do load may represent only 12.25% of the total number of files on the ANU intranet.

There is a relatively small increase in average loadable pages on a server when crawling from the server home page compared to crawling the entire ANU intranet. We suspect that expanding to an entire Web crawl would also yield an increase in the number of pages discovered on ANU intranet servers, but that this increase is likely to be small.

6.4 Mapping the Removed Dark Matter

No attempt has been made either to try to identify or to map removed ephemeral dark matter within the ANU intranet. However, we are able to report some figures for removed dead dark matter.

The search service we operate at ANU officially commenced operation on 29 July, 1999. We have kept lists of the URLs for each crawl of the intranet since that time. Figure 4 plots the growth in dead links for the longest time interval without policy changes in our crawler. The total number of pages downloaded on the 43rd day of the service was 124813. The numbers of pages which were no longer reachable on subsequent crawls up to the 79th day of the service are reported. The reason for the slight dip in dead links in the crawl on day 72 was probably due to server availability on that particular crawl. (Note that crawler page loading policy changes mean that we now load many more pages than we did when we commenced the service - the ANU intranet has not doubled in size in that time period.)

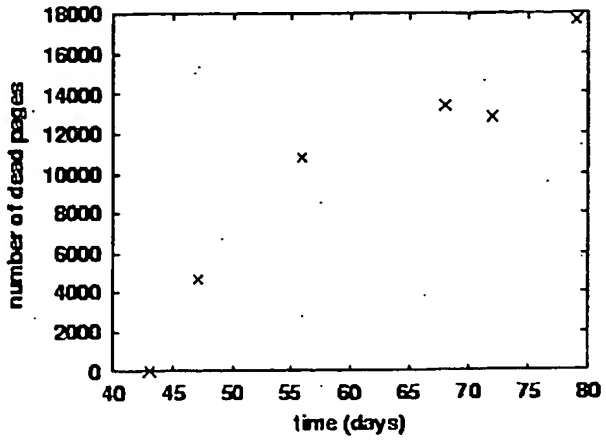

Figure 4: Increase in dead pages against time for ANU intranet crawls.

7 Size of the Publicly Indexable Web

As Lawrence and Giles's methodology [16] for estimating the size of the publicly indexable Web was based on crawling 2500 randomly selected Web servers from their home pages, it was an underestimate of the true size. There are often pages on a server which are not discovered from crawling from the home page. This is a straightforward consequence of the definition of lightness - pages which are reachable from the starting set of URLs. In their experiment, the starting set was just the single base URL of the server [14].

We conducted an experiment within the ANU intranet where we crawled 135 servers from their base URL and restricted our crawling to that server only. We then compared these numbers on each server with the number of pages on each server found when crawling the whole of the ANU (using the set of server URLs as our seed set). The average number of pages on a server when crawling was restricted to the server itself was 854. In contrast, when crawling all servers, the average number of pages on a server was 1331. Calculating the geometric mean of the ratios of increase on each server gives a mean ratio increase of 2.18. The reason this is so large (larger than the difference between the averages) is again due to the skew of data on the servers. The differences between the two methods of obtaining pages is plotted for each server (by decreasing server size) in figure 5.

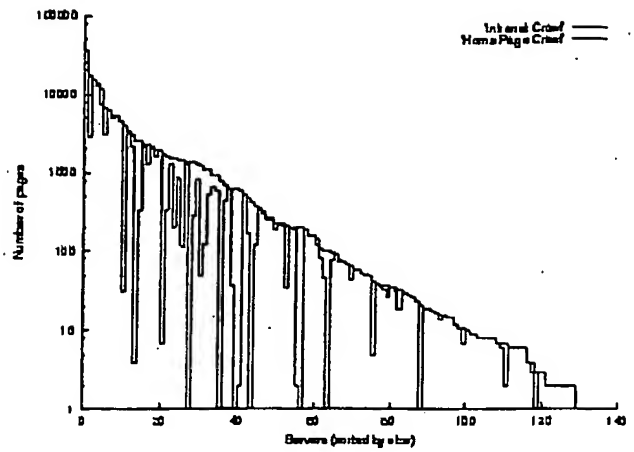

Figure 5: Histogram of the number of pages loaded on 135 servers when crawling the ANU intranet compared to crawling from their home pages only.

The difference between these figures and those given earlier based on our directory listings is due to the difference in the number of servers being crawled, and that we were reporting average percentages of pages on a server in different categories. The geometric mean of the ratios of increase for the servers with directory listings is 1.98, which is a little less than for the set of 135 servers. One inference which could drawn is that there is a greater degree of page connectedness within the servers for which we have directory listings.

Lawrence and Giles' estimate was that the Web contained 800 million pages in February 1999. If our results carry over to the Web at large (and we have no reason to believe that they would not), the size of the Web at that time, using the geometric mean ratio increase in server size, would have been 1744 million pages. We note that our estimate is an underestimate, in that if we could crawl a much larger subset of the Web, we might discover even more links to parts of the ANU which are not linked to within ANU.

8 Future Work

In future, we would like to expand the number of servers included in our directory listings. We are also interested in examining access logs for information about how often dark matter pages are actually downloaded. A similar study conducted on one or more intranets other than the ANU's would help to verify that our figures extend to the Web at large. However we have no reason to believe that the ANU is an unusual case.

9 Implications

As our study has shown, there is a large amount of dark matter on the Web. Most disturbingly for information providers, a significant percentage of it is reachable, but ignored by search engines. While some people may not care that the likelihood of their information ever being discovered is next to nil, most probably do.

9.1 Intentionally Public Information

Information on the Web has varying degrees of "public visibility". The intention of the information

provider about how public their information should be is partly captured by the mechanisms they use to prevent people gaining access to it.

The highest level of public visibility - unrestricted access - is achieved through two mechanisms: creating links to the information's URL from other public pages and publishing the URL to search engines. The intention of the information provider is that the information form part of the publicly indexable Web. (It also helps to advertise a URL in some fashion - emailing it to friends, posting it to newsgroups, or taking out full colour advertisements in colour supplements of major newspapers. However, these aspects are cultural, rather than systemic.) Lower levels of public visibility can be achieved by some of the different ways of restricting access to the information. For example, host or domain restrictions, `robots.txt` exclusions, user/password access, not creating links to the information; all these cause a restriction in access to smaller groups of potential users. These restriction mechanisms give rise to our categories of restricted and undiscovered private dark matter.

None of the other kinds of dark matter are intentionally dark. (Some of the ephemeral dark matter is not intended to be kept in posterity, but unless it is also restricted or private dark matter, there is nothing intentionally non-public about its content.) That this information *is* dark is cause for consideration about how best to make it light.

9.2 Decreasing the Dark Matter

The basic solution to limiting growth in dark matter is to address the issue of reachability. Making sure that intentionally public information is reachable is essential. Publishing the URLs to general search engines is clearly essential as well given that most information is discovered using search engines. Creating more public links to information is also very effective, but is difficult to accomplish other than on pages under the information provider's control.

In the longer term, specialised search engines (for example, the PubMed medical publications search service [17]) are likely to provide higher quality results since they are able to dedicate their resources to a single, albeit possibly broad, area of interest. General search engines will need to make greater use of distributed information retrieval techniques if they wish to gather results from a range of specialised search engines.

9.3 Policies of Information Publication

The existence of undiscovered dark matter on a Web document server may be deliberate or accidental. The information provider may wish for a page to be dark to all but a few individuals. If pages are deliberately kept dark, the information provider usually relies on some implicit understanding of how pages are made public on the particular server to ensure they remain dark. If a local search system is also provided on the server, there are two basic approaches to discovering the public information to be indexed: directory listings of designated non-private areas and crawling from designated public pages on the server.

In general, local search systems which use listings of all directories are more likely to uncover material that is not intended for publication. As such, they are best suited to organisations which have strong control over the creation and distribution of their information. For organisations where a more anarchic approach to information creation reigns, crawling from designated public pages on the server may be more appropriate. The latter approach is more widely understood by users due to its legacy as the default information discovery mechanism on the Web.

There are many reasons why sometimes it is essential to prevent information being indexed by either a local search server or a general public Web search server. Either of the two information discovery mechanisms given above can be adopted by document/search server administrators. The administrator of the server could make clear to all users the local publication policy so that they may deliberately keep their information dark. Similarly, making a policy clear may help users to prevent their information being accidentally kept dark.

10 Conclusions

The existence of dark matter on the Web is not a new phenomenon. Our experiments suggest that as little as 12.25% of all existing information on a server may be reachable to the majority of search engines. The remaining 87.75% of pages may thus be dark to many users. As much as 37% of all existing information on a server consists of loadable Web pages which remain as undiscovered dark matter. Exactly how much of this 37% is accidentally undiscovered dark matter rather than deliberately private dark matter is unknown. As the importance of the Web as the predominant information medium continues to rise, and the use of search engines continues to be a prime information discovery mechanism, the existence of such a huge quantity of information that is dark to these search engines assumes greater significance.

The dream of universal coverage of the Web by a single search engine seems to be long gone (with the notable exception of FAST [7]). Our categorisation of the different forms of Web matter indicate that not only was it unlikely to be feasible, but that it is impossible without infinite resources and omnipotent security privileges.

We believe that expanded use of meta-search systems (for example, MetaCrawler [8]) combined with increasing the provision of local search facilities on Web document servers is the most likely mechanism by which the quantity of dark matter, particularly undiscovered dark matter, may decline on the Web.

Acknowledgments

We wish to thank Helen Ashman and Steve Blackburn for providing comments on drafts of this paper. We also gratefully acknowledge the assistance of the various Web server administrators at the ANU who provided directory listings of the Web areas under their control.

The authors wish to acknowledge that this work was partly carried out within the Cooperative Research Centre for Advanced Computational Systems established under the Australian Government's Cooperative Research Centres Program. The core inspiration for this paper, the concept of dark matter on the Web, was first given to us by Paul Thistlewaite, our friend and colleague who died of cancer in February 1999.

Bibliography

1 Réka Albert, Hawoong Jeong, and Albert-László Barabási.

Diameter of the World-Wide Web.
Nature, 401(6749):130, Sept 1999.

2 AltaVista.
http://www.altavista.com/, 1999.

3 Krishna Bharat and Andrei Broder.
A technique for measuring the relative size and overlap of public web search engines.
In *Proceedings of the Seventh International World Wide Web Conference*, pages 379-388. 1998.

4 Diana Botluk.
Exposing the Invisible Web.
http://www.llrx.com/columns/exposing.htm, October 1999.

5 Sergey Brin and Lawrence Page.
The anatomy of a large-scale hypertextual Web search engine.
In *Proceedings of the Seventh International World Wide Web Conference*, pages 107-118, 1998.

6 K. Claffy, Tracie E. Monk, and Daniel McRobb.
Internet tomography.
Nature Web Matters, http://helix.nature.com, Jan 1999.

FAST Search.
http://www.alltheweb.com/, 1999.

8 Go2Net.
http://www.metacrawler.com/, 1999.

9 Google.
http://www.google.com/, 1999.

10 Graphic, Visualization, and Usability Center.
GVU's 10th WWW User Survey.
http://www.gvu.gatech.edu/user_surveys/survey-1998-10/,
October 1998.

11 B. A. Huberman and L. A. Adamic.
Evolutionary Dynamics of the World Wide Web.
http://www.parc.xerox.com/istl/groups/iea/www/growth.html
1999.

12 InvisibleWeb.
http://www.invisibleweb.com, 1999.

Martijn Koster.
A Standard for Robot Exclusion.
http://info.webcrawler.com/mak/projects/robots/norobots.h
1994.

14 Steve Lawrence.
private communication, 1999.

15 Steve Lawrence and C Lee Giles.
Searching the World Wide Web.
Science Magazine, 280(5360):98, April 1998.

16 Steve Lawrence and C Lee Giles.
Accessibility and Distribution of Information on the Web.
Nature, 400:107-109, 1999.

17 National Library of Medicine.
http://www4.ncbi.nlm.nih.gov/PubMed/, 1999.

18 Hrvoje Niksic.
GNU Wget Manual.
http://www.gnu.org/manual/wget/index.html, 1998.

NorthernLight.
http://www.northernlight.com/, 1999.

20 James E. Pitkow.
Summary of WWW characterizations.
In *Proceedings of the Seventh International World Wide Web Conference*, pages 551-558, 1998.

21 Danny Sullivan.
The AltaVista Size Controversy.
http://searchenginewatch.com/sereport/9707-avsize.html, 1997.

22 SUN Microsystems.
http://developer.java.sun.com/developer/, 1999.